Lasso 回归：从解释到预测

张沥今 [1]；魏夏琰 [2]；陆嘉琦 [2]；潘俊豪 [1*]

[1] 中山大学心理学系，广州，510006

[2] 浙江大学心理与行为科学系，杭州，310028

摘　要　传统的最小二乘回归法关注于对当前数据集的准确估计，容易导致模型的过拟合，影响模型结论的可重复性。随着方法学领域的发展，涌现出的新兴统计工具可以弥补传统方法的局限，从过度关注回归系数值的解释转向提升研究结果的预测能力也愈加成为心理学领域重要的发展趋势。Lasso 方法通过在模型估计中引入惩罚项的方式，可以获得更高的预测准确度和模型概化能力，同时也可以有效地处理过拟合和多重共线性问题，有助于心理学理论的构建和完善。

关键词　回归；正则化；预测；Lasso

1.　引言

心理学研究的目的在于“描述、解释、预测和影响行为”(彭运石, 李璜, 2011; Lippke & Ziegelmann, 2010)，探究变量间的关系是实现该目的必不可少的部分。回归分析作为一类评价变量间关系的方法，其思想已得到广泛推广，并且在各种主流统计分析软件中都可实现。回归分析是社会科学领域中最基础、最经典的定量分析方法(谢宇, 2010)，许多常见的统计检验(如，方差分析)也可以视作是线性回归模型的特例。回归模型的一般公式可以表示为：

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip} + \varepsilon_i \tag{1}$$

该模型包含 p 个预测变量，其中β_0为截距项，β_j表示第 j 个预测变量的回归系数($j = 1,2\ldots p$)，y_i表示第 i 个被试在结果变量上的观测值，x_{ij}表示第 i 个被试在第 j 个预测变量上的观测值，ε_i为残差项。

回归分析常被用于探索变量间的关系，同时也可以帮助研究者对结果变量进行预测。在采用回归模型分析数据的心理学研究中，最小二乘法(Ordinary Least Square, OLS)是最常用的模型系数估计方法(Helwig, 2017)。OLS 方法通过最小化结果变量的预测值与观测值之间的误差来估计回归模型中的参数，可以针对当前样本提供最准确的线性无偏估计(Chartterjee,

收稿日期：2019-12-13

* 国家自然科学基金项目(31871128)；教育部人文社会科学研究规划基金项目(18YJA190013)。

通讯作者：潘俊豪，E-mail: panjunh@mail.sysu.edu.cn

Hadi, & Price, 2000; Chartterjee & Hadi, 2006; Fomby, Hill, & Johnson, 1984; Maddala, 2002)。

但 OLS 方法关注于对当前数据集的无偏估计，容易导致模型发生过拟合现象(Yarkoni & Westfall, 2017)，即基于当前样本得到的回归模型结果在拟合同一总体的其他样本数据或用于预测未来观测数据时表现不佳，这一问题在预测变量较多，变量之间存在较高共线性或数据信噪比较低的情况下更为严重(Babyak, 2004; Helwig, 2017; McNeish, 2015)。过拟合的模型中往往会纳入不必要的冗余变量，并高估了部分预测因素的作用，削弱了模型的简约性(Babyak, 2004; Cohen, J., Cohen, P., West, & Aiken, 2003; Derksen & Keselman, 1992)。这些问题会对模型结论的推广和预测造成不可忽略的影响。

随着机器学习领域的蓬勃发展，涌现出了越来越多的统计工具用以弥补传统方法的局限。其中以 Lasso(Least absolute shrinkage and selection operator; Tibshirani, 1996)方法为代表的正则化(regularization)方法可以有效优化 OLS 估计、处理过拟合问题(Candes & Tao, 2007; Tibshirani, 1996; Tibshirani, Saunders, Rosset, Zhu, & Knight, 2005; Zou, 2006; Zou & Hastie, 2005)。正则化方法通过在模型估计中增加惩罚项的方式可以将过小的回归系数压缩到 0，以一定的估计偏差为代价从而获得更高的模型预测准确度和模型概化能力。该方法能够将冗余预测变量的估计系数压缩为 0，在压缩系数的同时起到变量筛选的作用，可以有效避免由于过拟合导致的模型概化能力不足的问题，获得更简约且具有较高预测效率的模型，有助于心理学理论的构建和完善。

Lasso 正则化方法自提出后吸引了诸多研究者的关注(Zou, Hastie, & Tibshirani, 2007)：由于该方法在变量筛选和模型稳定性上的出色表现，医学、经济学、神经科学等领域已有许多研究者采用 Lasso 方法建立模型进行预测(e.g., Fontanarosa & Dai, 2011; Lee, Chao, Ting, Chang, Huang, Wu, et al., 2014; Nguyen, Duong, Venkatesh, & Phung, 2015)。但是在神经科学以外的心理学领域中，对 Lasso 方法的运用却非常少(Johnson & Sinharay, 2011; McNeish, 2015; Yarkoni & Westfall, 2017)。其阻碍主要来自于对正则化等机器学习方法可解释性的质疑，这类方法常常不依赖于传统的假设检验，更多地采用数据驱动的方式进行探索和预测，因而被认为是一个“黑匣子”。吴喜之(2019)指出，事实上回归模型中单个回归系数同样不具备可解释性。例如，在回归模型的结果报告中，通常会出现这样的描述：“当保持其它预测变量不变时，该预测变量每变化一个单位，因变量变化 β 个单位”，但是这个前提条件几乎不可能成立。而除了研究结论的解释之外，模型的概化能力及预测能力同样值得关注。

在心理学研究中，以往由于受到计算机计算能力以及传统统计方法的限制，研究者在验证理论、检验变量间关系时，主要采用假设检验的方式。随着这类方法的普遍应用，其局限

日益突出，过拟合问题和可重复性危机也日益受到重视(胡传鹏 等, 2016; Nuzzo, 2014)。随着机器学习领域的蓬勃发展，新兴的数据科学工具已经在医疗健康等众多领域发挥出了巨大价值，在心理学领域，提升研究结论的预测能力将会成为未来重要的发展趋势(Yarkoni & Westfall, 2017)。

本文希望以 Lasso 方法为例，从理论出发，结合实例分析与具体应用现状，全面地为心理学研究者介绍 Lasso 回归的原理、实现步骤和优势，呼吁研究者在样本量较少或变量数目较多时采用更稳健的 Lasso 回归法来提升研究结论的可推广性。此外，本文还将介绍 Lasso 方法的多种扩展形式，及其在网络分析、潜变量建模中的应用。希望能够为研究者的实际应用提供参考，促进更多心理学研究者关注此类新兴的数据科学工具，以数据科学助力心理学的发展。

2. 传统方法及其局限

在标准的 OLS 回归中，回归模型的参数估计可以通过最小化损失函数得到，即最小化观察值与预测值之间的垂直平方距离，OLS 估计的损失函数公式具体如下(McNeish, 2015)：

$$L^{OLS}(\beta) = \|Y - X\beta\|^2 \quad (2)$$

其中L^{OLS}是损失函数，假定 n 为观察值个数，p 为预测变量个数(包括截距项)，X($n\times p$) 和 Y($n\times 1$)分别是预测变量矩阵和结果变量向量，β($p\times 1$)是回归系数向量。

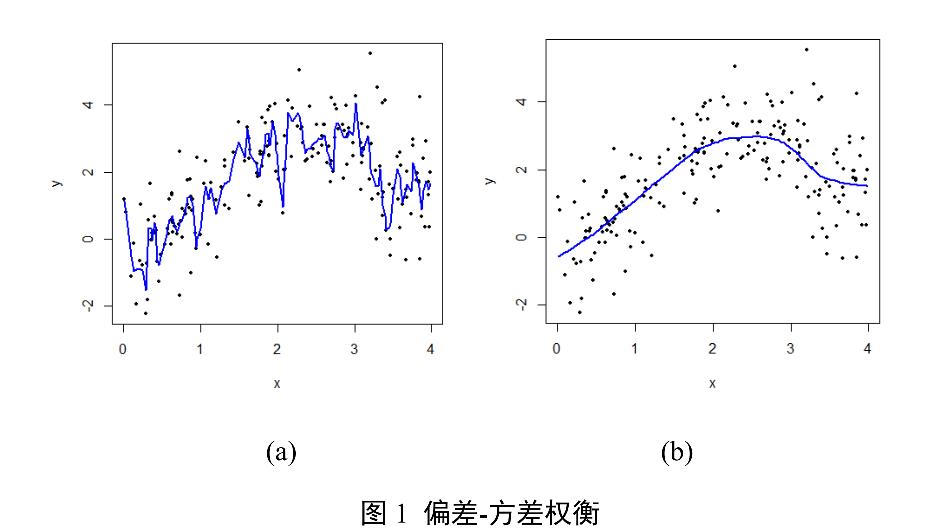
通过最小化 L^{OLS}，OLS 回归能够得到最好的线性无偏估计量 $\hat{\beta}^{OLS}$(Best Linear Unbiased Estimator, BLUE)，而且 OLS 估计的计算负担小，可以满足心理学领域的很多建模情境。但是，当研究中包含的预测变量数目较多时，OLS 估计法存在以下几点局限：

一是过度拟合(Overfit)，即建立的回归模型过于复杂，其中一些参数的显著性是由于抽样变异性(Sampling Variability)导致的，使得模型只适用于当前样本，缺乏概化能力(Generalizability)。

回归模型的预测误差可以被分解为偏差和方差两部分，其中偏差指预测值和真实值之间的差异，方差指预测值的离散情况。OLS 估计旨在通过控制估计偏差来降低模型的预测误差，但是参数的样本间方差会因此而增大，当前的参数估计结果可能仅适用于当前数据集，且估计结果易受到不同样本的微小波动的影响(如图 1a 所示，尽管模型对数据点的拟合较为准确，偏差较小，但这样的模型可能并不适用于其它样本)，容易出现过拟合现象。过拟合现象会导致模型在高估回归系数的同时低估其标准误，容易导致模型中无关联的冗余变量被

发现存在显著的预测作用，模型得到的结果可能仅适用于当前样本而无法推广到总体。当观察值个数 n 与预测变量个数 p 的比率越低时(即样本量不足)时，参数被错误解读的风险也越大(Babyak, 2004; Derksen & Keselman, 1992)。

相反地，如果使用当前数据集获得的参数估计存在着可以接受的偏差，那么参数的样本间方差会因有偏估计而减小，这样的估计结果反而具有更强的概化能力(如图 1b 所示)。因此，在实际数据分析中我们需要很好地处理这种偏差-方差权衡(Bias-Variance Tradeoff)问题。而传统的 OLS 估计关注对当前数据集的精确估计，在预测变量较多时不可避免地容易出现过拟合的估计结果，进而削弱模型的概化能力。



(a) (b)

图 1 偏差-方差权衡

二是多重共线性(Multicollinearity)，即在回归模型中多个预测变量间存在相关关系的现象，其中当预测变量间的相关系数为正负 1 时，即存在完全多重共线性。当模型存在较强的多重共线性时，OLS 估计得到的回归系数极易受到样本数据的微小波动的影响，估计的稳定性较差。回归系数的估计方差也会随着自变量间共线性的增强而增大(张凤莲，2010)。即当更换样本中的部分数据时，回归系数因为多重共线性的存在会产生较大的变化。这不仅会导致得到的回归模型缺乏概化能力，还会使某些重要变量的回归系数变得微不足道甚至与现实情况相反(Rao, 1976)。

此外，当模型存在较多的预测变量时，我们往往会采用逐步回归(Stepwise Regression)等方法增加或删减变量，以获得有效的预测变量集。但是该方法违背了回归分析推论的前提假设，即所有预测变量是作为整体固定存在的(Lockhart, Taylor, Tibshirani, R. J., & Tibshirani, R., 2014)，过度拟合带来的问题在使用逐步回归法进行模型选择时也会更加突出。此时用于统

计推断的 t 检验或 F 检验不仅无法遵循其适合的零假设分布，也无法拥有合适的自由度进行分析，基本的统计检验及其相关的 p 值将不适用于不断增减变量的模型选择。这种模型选择可能会使回归系数假设检验的一类错误率增大(Wilkinson, 1979)。

3. Lasso 方法

3.1 Lasso 方法介绍

相较于上文提及的 OLS 估计，正则化方法在 OLS 损失函数的基础上引入了惩罚函数，以惩罚过于复杂的模型。其具体公式可以表示为：

$$L^{Reg}(\beta) = L^{OLS}(\beta) + \lambda P(\beta) \tag{4}$$

其中，$L^{Reg}(\beta)$是惩罚后的损失函数，$L^{OLS}(\beta)$表示标准 OLS 损失函数，$P(\beta)$表示惩罚函数，$\lambda(\geq 0)$表示调整参数(Tuning Parameter)，用于控制回归系数压缩的程度，数值越大则惩罚力度越强。当$\lambda = 0$时，损失函数不对模型进行惩罚，$L^{Reg}(\beta)$即为 OLS 损失函数。而不同的惩罚函数$P(\beta)$则对应于不同的正则化方法。

Lasso 方法作为正则化方法的一种，它以回归系数的绝对值之和作为惩罚函数来压缩回归系数，即 $P^{\text{Lasso}}(\beta) = \lambda \sum_{j=1}^{p} |\beta_j|$，在参数估计中，由于绝对值符号难以进行拆解运算，可以将$|\beta_j|$转换为$\pm 1 * \beta_j$，其中+1 或-1 的具体符号与β_j的符号一致。即 Lasso 损失函数的公式可以表示为(McNeish, 2015)：

$$L^{\text{Lasso}}(\beta) = (|\mathrm{Y} - \mathrm{X}\beta|)^2 + \lambda \mathrm{W}^{\mathrm{T}}\beta \tag{5}$$

上述公式中，L^{Lasso}指 Lasso 回归模型的损失函数，$X(n \times p)$、$Y(n \times 1)$和 $\beta(p \times 1)$分别是预测变量矩阵、结果变量向量和回归系数向量，而 $W(p \times 1)$则是值为± 1 (符号与 β 向量中对应的数值一致)的向量。

相比其他正则化方法，如，Ridge 正则化采用回归系数的平方和为惩罚函数，对较小的回归系数估计值压缩力度更小，难以将冗余预测变量的系数压缩为 0，且对较重要的回归系数更容易进行过度压缩(Hesterberg, Choi, Meier, & Fraley, 2008)。Lasso 方法可以直接将冗余预测变量的回归系数压缩到 0 进而发挥变量选择的作用，获得精简且更有效率的预测变量集(Tibshirani, 1996)，同时也可以减少对重要回归系数的过度压缩。

Yarkoni 和 Westfall(2017)指出，相比于 OLS 估计法，Lasso 方法获得的模型通常能够更好地推广到新的数据集中。在 OLS 回归模型中，模型的 R^2(即结果变量的被解释率)通常会随着模型的复杂度增加。而 Lasso 方法不仅仅关注于解释当前的数据集(即得到更高的 R^2)，

也希望能够获得更简洁的模型以更好地推广到总体中。Lasso 方法从解释向预测的转变使得研究不仅仅指向于过去(即对当前数据集的解释)，同时也关注于未来(对新数据集的预测能力)。这一特性不仅有助于心理学理论的构建和完善，同时也可以一定程度上减少可重复性危机的影响。

此外，Lasso 方法也避免了在预测变量过多时采用 OLS 估计带来的过拟合和多重共线性的问题。而理论不完善且预测变量间存在共线性是心理学领域中较为常见的现象。当研究者的理论假设并不明确时，采用包含多重检验修正的验证性方法(如逐步回归)从理论上来说是错误的(Serang, Jacobucci, Brimhall, & Grimm, 2017)，后纳入的变量在这种情况下常常会因为与之前的变量存在相关而被削弱影响(Frank & Heiser, 2011)。Lasso 方法则将预测变量集视为整体，可以较好地应对这一问题。

由于惩罚项的引入，Lasso 方法在估计时所需要的计算量相对更高。Efron, Hastie, Johnstone 和 Tibshirani(2004)针对这一问题提出的最小角回归(Least Angle Regression, LARS)估计方法目前应用较为广泛。对于应用研究者来说，随着正则化方法的成熟，也已经发展出了可以直接进行 Lasso 回归建模的 R 软件包，对此本文将在下文详述。

3.2 Lasso 回归实现步骤

Lasso 回归建模通常包括参数 λ 的选择和 p 值的计算两部分，下文将详细介绍其方法原理，附录部分采用实例分析展示了如何在 R 软件中实现 Lasso 回归建模，并详细对比了 Lasso 回归和 OLS 回归方法。

3.2.1 参数λ的选择

参数 λ 的选择决定了回归系数被压缩的程度，不同的 λ 可能产生不同的结果。目前有以下两种常用的挑选参数λ最优值的方法(McNeish, 2015)：

第一种方法是机器学习领域的交叉验证(Cross-Validation)方法。具体过程如下：首先，将数据分成 K 个大小相同的样本，通常 K 可为 5、10 或 N(样本量)；然后选择 λ 的某个值，将前 K-1 份的数据采用 Lasso 方法估计模型，再将模型得到的回归系数用于第 K 份数据的验证，检验模型设立是否正确，并且将上述过程重复 K 次；最后，我们将得到某一 λ 值下模型的拟合值(如，线性模型的均方误差值)和标准误。交叉验证方法通常会重复上述过程 100 次，即选择 100 个不同的 λ 值，再以均标准误大小决定参数 λ 的取值。一般情况下，我们会选择均标准误最小时的 λ 值，但是有时选择 λ 的最小值意味着回归系数压缩幅度较小，

可能不能完全解决过拟合的问题。因此，有研究建议选择大于最小均标准误一个标准误时对应的参数 λ 值(Waldmann, Mészáros, Gredler, Fuerst, & Sölkner, 2013)。

另一种方法是信息标准(Information Criteria)，其参数 λ 的选择过程与交互验证基本相同，即针对多个不同的 λ 值，在每个 λ 值下，均采用 Lasso 方法拟合模型(使用全部数据)并计算得到信息标准的值(如，Akaike Information Criterion, AIC; Bayesian Information Criterion, BIC)。信息标准的具体计算公式分别如下所示：

$$\mathrm{AIC} = \mathrm{n}\log(RSS) + 2df \tag{7}$$

$$\mathrm{BIC} = \mathrm{n}\log(RSS) + df \tag{8}$$

其中 RSS 指的是残差平方和，*df* 则指自由度。通常我们会选择产生局部最小或整体最小的信息标准时参数 λ 的值(McNeish, 2015)。

目前大多研究者主要使用交叉验证方法来决定 λ 的数值(Obuchi & Kabashima, 2016)。

3.2.2 *p* 值的计算

大多变量选择的方法(如，逐步回归)得到的自由度或标准误是不正确的，这些方法在进行显著性检验时考察的不是应当作为整体存在的 k 个预测变量，而是经过筛选后的 m 个预测变量($m \leqslant k$, Thompson, 2001)。例如，对于样本量 n 为 101，k 为 50 的一个回归模型，F 检验的自由度[1]应为(50, 50)，但是如果逐步回归从 50 个预测变量中选出了五个预测变量，F 检验的自由度将变为(5, 95)。而据此计算得到的 p 值往往是不可靠的(Lockhart et al., 2014)。但是目前还没有较好的方法可以在不重复抽样或分割数据集的情况下处理 p 值的计算。

在 Lasso 回归中，对于没有完全压缩到零的回归系数，也难以计算其标准误并判断其显著性。鉴于此，Lockhart 等人(2014)提出了在不需重复抽样和分割数据的条件下，计算 Lasso 估计中 p 值的方法。该方法与传统的似然比检验相似。在标准的似然比检验中，我们需要计算全模型和限制模型(全模型中一些自由估计的参数在限制模型中被限制为 0)的偏差(偏差=-2log(似然值))，再通过卡方检验来比较嵌套模型间差异的显著性(限制模型嵌套于全模型)，进而进行模型选择。类似地，Lockhart 等人(2014)证明了结果变量(Y)的观察值和模型预测值($\mathrm{X}\hat{\beta}$)之间的协方差也可发挥类似上述似然比检验中“偏差”的作用，即在仅缺少某一预测变

---

[1] 在 OLS 回归分析中对多个回归系数进行单独或联合的显著性检验时可以用到 F 检验(张厚粲, 徐建平, 2015)，F 检验的自由度为(df_R，df_E)，df_R和df_E分别是回归平方和及残差平方和对应的自由度。对于一个有 n 个观测值和 k 个预测变量的回归方程，$df_R = k$，$df_E = n\text{-}1\text{-}k$。

量的模型中(即限制模型，该预测变量的回归系数被限制为 0)，加入该预测变量后(即全模型，所有预测变量的回归系数被自由估计)，计算模型协方差的变化值，再进行显著性检验就可以实现变量选择。这种方法在检验每一个预测变量的显著性时都纳入了其余所有预测变量的影响，避免了逐步回归中依次纳入变量时先纳入的变量对后纳入变量的影响。也不需要通过分离数据或重复抽样来进行推断性检验，操作相对简便。

为了演示 Lasso 回归的实现步骤和报告标准，附录部分采用实证数据详细展示了 Lasso 回归在 R 软件中的实现过程。分析采用 glmnet 软件包(Friedman, Hastie, & Tibshirani, 2010)进行参数 λ 的选择，采用 covTest 软件包(Lockhart et al., 2014)计算参数估计的 p 值。

4. Lasso 回归的应用

Lasso 回归的优点主要体现在它能够充当稳定的变量筛选器、建立更具有概化能力和预测能力的模型。在理论相对不够完善的研究中，研究者更加需要采用这类方法避免对当前样本的过度解释，探索适用于总体的规律。这种从解释到预测的转变有助于提升这类研究的理论意义和应用价值。

Lasso 方法的优良特性使得其在教育学、临床心理学、发展心理学等领域都有着广阔的应用前景。但心理学领域的研究中，只有少量研究采用了 Lasso 方法(如, Hartmann, Zeeck, & Barrett, 2010; Scheidt et al., 2012; Schmid, Taylor, Foldi, Berres, & Monsch, 2013)。McNeish(2015)也指出统计方法在心理学中的应用现状与统计学研究进展之间存在着较大的脱节。从统计学研究成果到心理学中的广泛应用往往都需要较长时间，这导致应用领域不能迅速地从统计学的最新研究中获益。基于此，下文将列举 Lasso 方法在临床心理学和认知神经科学领域的实际应用来展现该方法的具体使用与优势，希望能够为研究者们使用 Lasso 方法提供参考。

4.1 Lasso 在认知神经科学中的应用

在神经科学领域，Lasso 已被成功应用在全基因组关联研究(Genome Wide Association Study, GWAS)或候选基因研究中筛选基因位点(Single Nucleotide Polymorphism, SNPs; Ayers & Cordell, 2010; Shi et al., 2011)、检测基因与基因之间的交互作用(D’Angelo, Rao, & Gu, 2009; Li, Das, Fu, Li, R., & Wu, 2011)、以及根据 GWAS 结果进行风险预测(Kooperberg, LeBlanc, & Obenchain, 2010)。全基因组关联研究能够发现影响神经和精神疾病的风险基因，在进行 GWAS 研究时，往往会涉及大量的基因位点。此类涉及大量变量的基因研究往往存在研究结

果难以重复的问题(Kohannim et al., 2012)。采用 Lasso 方法，能够恰当减少 SNPs 的数量，筛选出与结果变量稳定相关的基因，建立可重复的模型。另外，传统的 GWAS 分析将每个基因的作用看成是独立的，忽略了它们之间可能存在连锁不平衡结构(Linkage Disequilibrium, LD)，即部分变异更可能被一起遗传。综上，在基因分析中采用 Lasso 方法的优势主要有(Cho, Kim, Oh, Kim, & Park, 2009; Cho et al., 2010; Lin et al., 2009; Malo, Libiger, & Schork, 2008; Shi et al., 2011)：(1)能够处理基因组的多维度问题；(2)能够处理由于 LD 引起的多重共线性问题；(3)能够处理多重比较的问题。

Kohannim 等人(2012)为了减少相关基因数量，筛选出与大脑结构具有可靠相关的基因，采用 Lasso 回归来检测哪些基因能够影响颞叶体积(神经退行性疾病的生物标志)。研究收集了 729 名老年被试的全基因组数据以及相关的协变量数据，结果变量为被试的颞叶体积测量。通过 Lasso 回归从备选 SNPs 中筛选出对结果变量影响最有效的一组 SNPs。最终得到了 22 个显著影响颞叶体积的基因。随后，为了检验基因结果的可重复性，他们在另一批独立的健康青年群体身上针对相关性最高的 MACROD2 基因进行了重复验证。在这批独立的青年群体身上同样发现了 MACROD2 基因对于大脑结构存在影响，验证了通过 Lasso 回归分析得到的基因相关结果的稳健性。

4.2 Lasso 在临床心理学中的应用

由于临床样本收集的困难以及研究者们对众多心理疾病的认识不够清晰，临床研究中往往会考虑较多变量的影响，导致观测值数量与预测因子数量的比值较小(Demjaha et al., 2017)。另外，临床评估要求我们建立能够进行稳定推断的模型。此时如果使用传统的逐步回归方法来进行变量筛选，容易出现过拟合问题。而使用 Lasso 方法能够获得稳定的参数估计并提高预测准确性(Harrell, 2015)，更加符合临床评估的要求。

基于此，Demjaha 等人(2017)调查影响首发性精神病抗治疗性(Treatment Resistance)的因素时，追踪了 323 名患有首发性精神病的患者，采用 Lasso 多元回归方法分析耐药性与临床及人口学变量之间的相关。Lasso 多元回归分析结果显示诊断为精神分裂症、阴性症状、首次发病年龄小、较长时间未接受精神病治疗这几个因素能显著预测被试的精神病抗治疗性。

另外，在患病早期识别患者是采取有效临床干预与治疗的先决条件，Lasso 方法已被成功应用于识别潜在的患者。Schmid, Taylor, Foldi, Berres 和 Monsch(2013)对 29 个后来发展为阿尔兹海默症的患者以及相应条件匹配的 29 个对照正常人进行了为期八年的追踪，调查了被试的客观行为测量以及神经心理学的功能变化情况。由于研究变量(k=115)相对于观测值

(n=29)来说数量过大，采用一般的回归方法易导致严重的过拟合问题。为了获得更具有预测能力的模型，研究者采用 Lasso 回归来识别哪些变量能够在早期区分未来将发展成阿尔兹海默症的人群与正常对照人群。最终从 115 个预测变量中筛选出了 11 个最具预测力的变量，能够有效地在早期区分两类人群。

5. Lasso 的扩展

5.1 Lasso 的扩展形式

在 Lasso 的基础上，研究者根据回归分析中自变量的不同特性，采用不同形式的惩罚函数，建立和发展出了多种正则化模型，例如松弛 Lasso (Relaxed Lasso; Meinshausen, 2007)，自适应 Lasso (Adaptive Lasso; Zou, 2006) ，Bayesian Lasso(Park & Casella, 2008)，Fused Lasso (Tibshirani et al., 2005)和 Group Lasso (Yuan & Lin, 2006)等。下文将介绍几种 Lasso 扩展形式的原理和对应的 R 语言软件包。

5.1.1 松弛 Lasso

当观测指标数p远大于观测样本量N时，Lasso 方法的收敛速度较慢(Fan & Peng, 2004)。由于 Lasso 方法无法同时在计算复杂度与收敛速度上达到令人满意的折中，Meinshausen(2007)在 Lasso 的基础上提出了一个两阶段分析方法——松弛 Lasso (Relaxed Lasso)。在松弛 Lasso 的分析中，模型选择和参数估计被分割成两个独立的过程。该方法首先采用普通的 Lasso 回归筛选出合适的预测变量，第二步再对筛选出的变量进行系数估计。此时会通过调整参数 Φ 改变惩罚力度($\lambda_2 = \Phi * \lambda$, $1 > \Phi \geq 0$, λ, λ_2分别为第一、二步估计中采用的调整参数)，削弱或消除惩罚项的作用来减小变量的系数估计偏差。当 Φ=1 时，系数估计值与普通 Lasso 方法得到的估计值一致；当 Φ=0 时，此时系数估计值与 OLS 方法的估计值相同。松弛 Lasso 在兼顾计算复杂度的同时拥有比 Lasso 更快的收敛速度(Meinshausen, 2007)。理论和数值结果已表明，对于高维数据，松弛 Lasso 能够产生更稀疏的模型以及与 Lasso 相等或更小的预测损失。

对于松弛 Lasso 的应用，已有较为完备的软件包可供使用。R 语言中的 relaxo 包(Meinshausen, 2019)是专门用于进行松弛 Lasso 分析的软件包，仅需要调用 cvrelaxo 或 relaxo 函数即可非常便捷地获得松弛 Lasso 的解。本文也采用实证数据进行了松弛 Lasso 回归的演示（第二步中参数 Φ 被固定为 0，即采用 OLS 回归法），并对比了传统的 OLS 回归估计的

结果。发现松弛 Lasso 回归仅采用两个预测变量就基本达到了 OLS 回归采用 5 个变量所获得的预测能力。

5.1.2 自适应 Lasso

Lasso 方法通过调整参数 λ 来控制回归系数的压缩程度(Tibshirani, 1996)。当研究者通过交叉验证方法选择并设定 λ 为某个固定的值时，Lasso 方法会对所有变量施加相同程度的惩罚，尽管这相比于 Ridge 正则化方法已经一定程度上减少了对重要回归系数的过度压缩，但仍然不可避免地可能会对重要变量的系数进行压缩，产生一定的估计偏差(Fan & Li, 2001)。Zou (2006)通过在惩罚项前增加自适应权重对 Lasso 算法进行了改进，提出了自适应 Lasso 方法(Adaptive Lasso)。在自适应 Lasso 方法中，任选一个 $\gamma>0$，则权重向量 $\hat{w}=\frac{1}{|\hat{\beta}|^{\gamma}}$，此处可以采用 OLS 方法得到的系数估计值作为初始系数估计值 $\hat{\beta}$，则自适应 Lasso 中的惩罚项可以表示为：

$$P^{\text{aLasso}}(\beta)=\lambda\sum_{j=1}^{p}\hat{w}_j|\beta_j|$$

自适应 Lasso 中的自适应权重系数依赖于数据，不同变量的回归系数受到的惩罚程度不同。对于初始系数估计值较大的变量，其权重系数较小，从而会受到更小的惩罚。而初始估计值较小的变量对应较大的权重系数与较大的惩罚。因此，采用自适应 Lasso 进行变量选择能够使得重要的变量更易进入模型，而不重要的变量更易被剔除，在更好地实现变量选择的同时也能够有效减小系数估计的偏差。相比于 Lasso 方法，自适应 Lasso 方法也更适用于观测指标数 p 和样本量 N 的比值非常大的情况。目前，R 语言中的 glmnet (Tibshirani et al, 2019)、msgps(Hirose, 2019)以及 parcor (Kraemer & Schaefer, 2019)等软件包均能进行自适应 Lasso 分析。另外，SAS 软件中的 Proc GlmSelect 也能实现自适应 Lasso 的分析。

5.1.3 贝叶斯 Lasso

在频率学派中， Lasso 方法通过在似然函数值增加惩罚项的方式来减少模型参数，实现正则化。而在贝叶斯方法中，如果选择了合适的先验分布，先验分布的对数形式就会扮演惩罚项的角色。例如，Tibshirani (1996)认为在贝叶斯方法下如果对参数 θ_j 提供同样的、相互独立的双指数先验分布$\frac{\lambda}{2}exp(-\lambda|\theta_j|)$，就可以实现 Lasso 正则化。双指数先验分布与零均值正态分布一样具有单峰性和对称性，但其峰度比正态分布更大。其中， λ 值越大，概率密度函

数越集中在零附近。

此外，频率学派中能够实现 Lasso 方法的算法(如，Efron et al., 2004; Friedman et al., 2010; Wu & Lange, 2008)并不能提供有效的标准误估计，这对于频率学领域中 Lasso 方法的应用造成了阻碍(Kyung, Gill, Ghosh, & Casella, 2010)。而贝叶斯 Lasso 可以通过 Gibbs 采样法提供有效的标准误估计(Kyung et al., 2010)。Park 和 Casella (2008) 以及 Hans (2009) 提出的贝叶斯 Lasso 回归模型也能够在估计未知系数的同时估计正则化参数，避免了使用传统交叉验证方法所需的大量计算负担，有着非常广阔的应用前景。而应用研究者也已经可以采用 R 语言中的 blasso 软件包(Gramacy, 2019)非常方便的进行贝叶斯 Lasso 回归建模。

5.2 Lasso 的扩展应用

在回归模型中，Lasso 方法还可以被用于筛选中介变量(Serang et al., 2017)；而在回归模型之外，正则化方法也逐渐被应用于结构方程模型(Structural Equation Modeling)和心理网络模型(Psychological Network Models; Epskamp, Borsboom, & Fried, 2018)中。

5.2.1 潜变量模型

潜变量模型主要被用于分析问卷测量的数据，它在模型估计时考虑了测量误差的影响。在潜变量建模领域，正则化方法已经引起了方法学家的重视，逐渐被引入到结构方程建模分析中，如，采用贝叶斯 Ridge 正则化或 Lasso 正则化方法解决传统的验证性因子分析限制过于严格的问题(Muthén & Asparouhov, 2012; Pan, Ip, & Dubé, 2017)，在 MIMIC 模型(Multiple Indicators and Multiple Causes, MIMIC)中利用正则化方法进行预测变量的筛选(Jacobucci, Brandmaier, & Kievit, in press)等。

目前最为流行的潜变量分析软件 M*plus* (Muthén, L, K., & Muthén, B, O., 1998-2019)已经可以采用 Ridge 正则化方法进行结构方程建模，其应用也十分普遍(张沥今，陆嘉琦，魏夏琰，潘俊豪, 2019)。也有专门的 R 语言软件包“blcfa”(Pan, Zhang & Ip, 2019)可以进行贝叶斯 Lasso 验证性因子分析，以及“regsem”软件包(Jacobucci, 2019)可以帮助研究者利用 Ridge 正则化或 Lasso 正则化方法进行探索性因子分析、建立 MIMIC 模型等。遗憾的是，由于 Lasso 方法与潜变量模型结合的方法在近两年才得到发展，目前尚未得到普遍应用。

5.2.2 网络模型

心理网络模型(Psychological Network Models)采用节点(Nodes)代表可观测变量，边

(Edges)代表可观测变量间的联系，边的权重代表变量间联系的强度。这种模型认为某些心理过程、状态(如，认知过程，精神病理症状)是同时发生的，因此它关注各个可观测变量在网络中的相互作用。心理网络模型可以帮助研究者深入了解可观测变量间的关系，是潜变量模型的有力补充。近年来，心理网络模型被广泛应用于人格心理学和临床心理学等研究领域(如, Costantini et al., 2019; Richetin, Preti, Costantini, & De Panfilis, 2017)。

由于该模型考察的变量和参数数目较多，为了避免过拟合问题、降低一类错误率，研究者在采用网络分析时通常都会结合 Lasso 方法进行变量筛选。自适应 Lasso 和图 Lasso(Graphical Lasso)等方法都可以帮助研究者获得稀疏的、具有更强概化能力的网络模型，如：Marcus, Preszler 和 Zeigler-Hill(2017)使用自适应 Lasso 方法建立了黑暗人格(Dark Personality)网络模型；Costantini 等人(2015a)基于自适应 Lasso 网络模型发展了责任感变量的内隐测量工具； Di Pierro, Costantini, Benzi, Madeddu 和 Preti (2018)则使用图 Lasso 方法建立自恋特质的精神病理学网络模型。这类网络模型可以通过 qgraph (网络分析软件包; Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012)和 glasso 软件包(图 Lasso 软件包; Friedman, Hastie, & Tibshirani, 2019)实现；为了方便应用研究者使用这类方法，Costantini 等人(2015b, 2019)详细阐述了自适应 Lasso 网络模型及图 Lasso 网络分析模型的原理及其在 R 软件中的实现方法。

6. 讨论

6.1 应用建议

在心理学研究中，研究者们常常主要关注于对变量间关系的解释，但是 Yarkoni 和 Westfall(2017)指出这种视角导致大量的心理学研究虽然探究了关系复杂的心理机制，但是这些模型却很难准确地预测未来的行为。随着可重复性问题日益受到重视，如何利用统计方法、规范研究流程来提供可重复性危机的解决方案也逐渐成为心理学领域的热点问题(Giordano & Waller, 2019; 胡传鹏 等，2016; Spellman, 2015)。过度关注对当前数据集的解释带来的过拟合现象也是造成可重复性危机的关键问题，针对该问题，研究者已经提出了一系列应对措施。例如，根据检验力和效应量在实验前计算样本量，将 p 值临界值修改为 0.005 的同时提高样本量以降低二类错误率(Benjamin et al., 2018)。但有些研究(如，临床研究)难以收集到足够的样本量，且在理论不够完善的情况下变量数目较多也是非常常见的现象，而过拟合问题在这种情况下会更为严重(Babyak, 2004; McNeish, 2015)。

因此，有研究者指出新的统计分析工具(如，正则化方法、贝叶斯方法)有望避免假设检验的局限，降低可重复性危机(胡传鹏 等，2016; Benjamin et al., 2018)。也有越来越多的研究者指出机器学习领域的工具有希望帮助心理学成为一门更有预见性的科学，且从解释向预测的转变或许可以帮助研究者更好地理解行为及其背后的机制(Rosenberg, Casey, & Holmes, 2018; Serang et al., 2017)。

以 Lasso 为代表的正则化模型在机器学习领域发挥着越来越重要的作用，目前也已经广泛应用于生物医学等领域，在心理学领域中正则化稀疏模型也可以帮助研究者进行变量筛选，解决模型中的过拟合问题，控制一类错误率等 (刘建伟, 崔立鹏, 刘泽宇, 罗雄麟, 2015; 许树红, 王慧, 孙红卫, 王彤, 2017)。在小样本及变量数目较多的情况下，Lasso 方法都有着更优良的表现，也越来越多地被应用于心理学领域，在临床心理学和认知神经科学之外，Lasso 回归在教育心理学、人格心理学等领域中也可以发挥其价值。因此，本文希望通过对 Lasso 回归方法原理和应用的介绍，展现正则化模型的价值，进而促进机器学习领域的工具在心理学领域发挥更大的作用。同时，我们也呼吁应用研究者在变量数目较多或样本量不足的时候采用 Lasso 方法进行建模分析。

6.2 Lasso 方法的局限和展望

阻碍 Lasso 回归应用的主要问题是其难以获得标准误估计值。这一方面会影响 p 值的计算，对此 Lockhart 等人(2014)提出的方法及对应的 R 软件包可以有效地弥补这一问题，但我们同时也希望研究者在应用这类机器学习方法时能够跳出显著性检验思维，更多地关注模型整体的预测能力。另一方面，无法获得标准误也会影响效应量和置信区间的计算，但 Lasso 方法的扩展形式贝叶斯 Lasso 可以有效地进行标准误、可信区间(Credible Interval)的估计，弥补了这一问题。随着贝叶斯统计的流行(Van de Schoot et al., 2017)，这种贝叶斯 Lasso 方法未来也有望得到更深入的发展和应用。

此外，主流的许多统计软件都无法实现 Lasso 回归方法(如，SPSS，M*plus*)，这极大地阻碍了 Lasso 方法的应用。在 R 软件中能够实现 Lasso 方法的软件包虽然多，却也各有各的局限。Rstudio 的首席科学家、ggplot2 软件包的作者 Hadley Wickham 在采访(邱怡轩, 2019)中也提到，他在课上会建议学生尝试一些更为稳健的回归方法，如 Lasso 类的统计方法。但他指出目前有大概 13 个关于 Lasso 方法的 R 包，但是每一个都不够完善，如，不能处理缺失值、分类变量等等，因此他计划将整合这些软件包以制作一个更高效的分析工具。相信随着正则化模型及其配套分析工具的成熟，应用研究者也可以更便捷地采用正则

化方法进行建模分析。

最后，Lasso 方法在回归模型之外的应用才刚刚起步，而 Lasso 方法的优良特性也使得其在处理复杂模型(如，潜交互模型、密集追踪模型等)时更具潜力。希望随着 Lasso 方法的发展，方法学家也能够在各个领域充分发挥 Lasso 方法的价值。未来研究也需要进一步对比 Lasso 方法与其它正则化方法并探索其分别适用的建模场景，为应用研究提供建议。

参考文献

胡传鹏, 王非, 过继成思, 宋梦迪, 隋洁, 彭凯平. (2016). 心理学研究中的可重复性问题: 从危机到契机. *心理科学进展, 24*(9), 1504–1518.

刘建伟, 崔立鹏, 刘泽宇, 罗雄麟. (2015). 正则化稀疏模型. *计算机学报, 38*(7), 1307–1325.

彭运石, 李璜. (2011, 十月). *论西方心理学发展中的说明与理解之争.* 文章展示于第十四届全国心理学学术会议, 北京.

邱怡轩, *统计之都访谈第 9 期: Hadley Wickham.* 2019-8-30 取自 https://mp.weixin.qq.com/s/IPejDdwIFIx93UxsRwtQ1Q

吴喜之. (2019). *从模型驱动的集体推断到数据驱动的个体预测.* 第 12 届中国 R 语言会议, 北京.

谢宇. (2010). *回归分析.* 北京：社会科学文献出版社.

许树红, 王慧, 孙红卫, 王彤. (2017). 基于 lasso 类方法的 I 类错误的控制. *中国卫生统计, 4,* 660–667.

张厚粲, 徐建平. (2015). *现代心理与教育统计学.* 北京：北京师范大学出版社.

张凤莲. (2010). *多元线性回归中多重共线性问题的解决办法探讨*(硕士学位论文). 华南理工大学, 广州.

张沥今, 陆嘉琦, 魏夏琰, 潘俊豪. (2019). 贝叶斯结构方程模型及其研究现状. *心理科学进展. 27*(11), 1812–1825.

Ayers, K. L., & Cordell, H. J. (2010). SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genetic Epidemiology, 34*, 879–891.

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine*, *66*, 411–421.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B., Wagenmakers, E. J., Berk, R., & Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour, 2*, 6–10.

Candes, E., & Tao, T. (2007). The dantzig selector statistical estimation when p is much larger than n. *Annals of Statistics, 35,* 2313–2351.

Chartterjee, S., Hadi, A. S., & Price, B. (2000). *Regression by Example: 3rd Edition.* Hoboken: John Wiley and Sons.

Chartterjee, S. & Hadi, A. S. (2006). *Regression by Example: 4th Edition.* Hoboken: John Wiley and Sons.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Mottus, R., Waldorp, L. J., & Cramer, A. O. J. (2015a). State of the aRt personality research: A tutorial on network analysis of personality data in R. *Journal of Research in Personality, 54*, 13–29.

Costantini, G., Richetin, J., Borsboom, D., Fried, E., Rhemtulla, M., & Perugini, M. (2015b). Development of Indirect Measures of Conscientiousness: Combining a Facets Approach and Network Analysis. *European Journal of Personality, 29*, 548–567.

Costantini, G., Richetin, J., Preti, E., Casini, E., Epskamp, S., & Perugini, M. (2019). Stability and variability of personality networks. A tutorial on recent developments in network psychometrics. *Personality and Individual Differences, 136*, 68–78.

Cortez, P., & Silva, A. (2008, April). Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds. *Proceedings of 5th FUture BUsiness TEChnology Conference* (pp. 5–12). Porto, Portugal.

Cho, S., Kim, H., Oh, S., Kim, K., & Park, T. (2009). Elastic-net regularization approaches for genome wide association studies of rheumatoid arthritis. *BioMed Central Procedings. 3*(Suppl.), S7–S25.

Cho, S., Kim, K., Kim, Y. J., Lee, J. K., Cho, Y. S., Lee, J. Y., et al. (2010). Joint identification of multiple genetic variants via elastic net variable selection in a genome wide association analysis. *American Journal of Human Genetics, 74,* 416–428.

D'Angelo, G. M., Rao, D., & Gu, C. C. (2009). Combining least absolute shrinkage and selection operator (LASSO) and principal-components analysis for detection of gene-gene interactions in genome-wide association studies. *BioMed Central Procedings, 3*(Suppl.), S7–S62.

Di Pierro, R., Costantini, G., Benzi, I. M. A., Madeddu, F., & Preti, E. (2018). Grandiose and entitled, but still fragile: A network analysis of pathological narcissistic traits. *Personality and Individual Differences, 140*, 15–20.

Demjaha, A., Lappin, J. M. , Stahl, D. , Patel, M. X. , Maccabe, J. H. , & Howes, O. D. , et al. (2017). Antipsychotic treatment resistance in first-episode psychosis: Prevalence, subtypes and predictors. *Psychological Medicine, 47*(11), 1–9.

Derksen, S., & Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, *45*, 265–282.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*, 407–499.

Fomby, T. B., Hill, R. C., & Johnson, S. R. (1984). *Advance Econometric Methods.* New York, Berlin, Heidelberg, London, Paris, Tokyo: Springer-Verlag.

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial

paper. *Behavior Research Methods, 50*(1), 195–212.

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network visualization of relationships in psychometric data. *Journal of Statistical Software, 48*(4), 1018.

Fan, J., & Li, R. (2001). "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association, 96*, 1348–1360.

Fan, J., & Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics, 32*, 928–961.

Fontanarosa, J. B., & Dai, Y. (2011). Using lasso regression to detect predictive aggregate effects in genetic studies. *BioMed Central Procedings, 5*(Suppl.9), 69–74.

Frank, L. E., & Heiser, W. J. (2011). Feature selection in feature network models: Finding predictive subsets of features with the positive lasso. *British Journal of Mathematical & Statistical Psychology, 61*(1), 1–27.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*, 1–22.

Friedman, J., Hastie, T. & Tibshirani, R. *Bayesian Lasso/NG, Horseshoe, and Ridge Regression.* Retrieved August 30, 2019, from https://www.rdocumentation.org/packages/monomvn/versions/1.9-10/topics/blasso

Giordano, C., & Waller, N, G. (2019). A Neglected Aspect of the Reproducibility Crisis: Factor Analytic Monte Carlo Studies, *Multivariate Behavioral Research, 55*(1), 152.

Hans, C. (2009). Bayesian Lasso regression. *Biometrika, 96*(4), 835–845.

Harrell, F. E., Jr. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis, 2nd.* New York: Springer-Verlag.

Hartmann, A., Zeeck, A., & Barrett, M. S. (2010). Interpersonal problems in eating disorders. *International Journal of Eating Disorders, 43*(7), 619–627.

Helwig, N. E. (2017). Adding bias to reduce variance in psychological results: A tutorial on penalized regression. *The Quantitative Methods for Psychology, 13*(1), 1–19.

Hesterberg, T., Choi, N. H., Meier, L., & Fraley, C. (2008). Least angle and $\ell 1$ penalized regression: A review. *Statistics Surveys*, *2*, 61–93.

Hirose, K. Retrieved August 19, 2019, from https://www.rdocumentation.org/packages/msgps/versions/1.3.1

Jacobucci, R. *regsem: regularized structural equation models. R package version 1.3.9.* Retrieved June 01, 2019, from https://cran.r-project.org/web/packages/regsem/index.html

Jacobucci, R., Brandmaier, A., & Kievit, R. (accepted). A practical guide to variable selection in structural equation

models with regularized MIMIC models. *Advances in Methods and Practices in Psychological Science.*

Johnson, M., & Sinharay, S. (2011). Remarks from the new editors. *Journal of Educational and Behavioral Statistics, 36*, 3–5.

Kohannim, O., Hibar, D. P., Stein, J. L., Jahanshad, N., Hua, X., & Rajagopalan, P., et al. (2012). Discovery and replication of gene influences on brain structure using lasso regression. *Frontiers in Neuroscience, 6*, 1–13.

Kooperberg, C., LeBlanc, M., & Obenchain, V. (2010). Risk prediction using genome-wide association studies. *Genetic Epidemiology, 34*(8), 643–652.

Kyung, M., Gill, J., Ghosh, M., & Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis, 5*(2), 369–411.

Kraemer, N., & Schaefer, J. *parcor: Regularized estimation of partial correlation matrices.* Retrieved Sept ember 04, 2019, from https://www.rdocumentation.org/packages/parcor/versions/0.2-6

Lee, T. F., Chao, P. J., Ting, H. M., Chang, L., Huang, Y. J., Wu, J. M., et al. (2014). Using Multivariate Regression Model with Least Absolute Shrinkage and Selection Operator (LASSO) to Predict the Incidence of Xerostomia after Intensity-Modulated Radiotherapy for Head and Neck Cancer. *PLoS ONE, 9*(2), e89700.

Li, J., Das, K., Fu, G., Li, R., & Wu, R. (2011). The Bayesian lasso for genome-wide association studies. *Bioinformatic*s, *27*, 516–523.

Lin, Y., Zhang, M., Wang, L., Pungpapong, V., Fleet, J. C., & Zhang, D. (2009). Simultaneous genome-wide association studies of anti-cyclic citrullinated peptide in rheumatoid arthritis using penalized orthogonal-components regression. *BioMed Central Procedings*, *3*(Suppl), S17–S20.

Lippke, S., & Ziegelmann, J. P. (2010). Theory-based health behavior change: Developing, testing, and applying theories for evidence-based interventions. *Applied Psychology, 57*(4), 698–716.

Lockhart, R., Taylor, J., Tibshirani, R. J., & Tibshirani, R. (2014). A significance test for the lasso. *The Annals of Statistics*, *42*, 413–468.

Maddala, G. S. (2002). *Introduction to Econometrics: 3rd Edition.* John Willey and Sons Limited, England.

Malo, N., Libiger, O., & Schork, N. J. (2008). Accommodating linkage disequilibrium in genetic-association analysis via ridge regression. *American Journal of Human Genetics, 82*, 375–385.

Marcus, D. K., Preszler, J., & Zeigler-Hill, V. (2017). A network of dark personality traits: What lies at the heart of darkness? *Journal of Research in Personality, 73*, 56–62.

Mcneish, D. M. (2015). Using Lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences. *Multivariate Behavioral Research, 50*(5), 471–484.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics and Data Analysis, 52*(1), 374–393.

Meinshausen, N. *Relaxed Lasso.* Retrieved June 01, 2019, from https://www.rdocumentation.org/packages/relaxo/versions/0.1–2

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17*(3), 313–335.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's Guide. Eighth Edition.* Los Angeles, CA: Muthén & Muthén.

Nguyen, T., Duong, T., Venkatesh, S., & Phung, D. (2015). Autism blogs: Expressed emotion, language styles and concerns in personal and community settings. *IEEE Transactions on Affective Computing, 6*(3), 312–323.

Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature. 506*(7487), 150–152.

Obuchi, T., & Kabashima, Y. (2016). Cross validation in lasso and its acceleration. *Journal of Statistical Mechanics: Theory and Experiment, 2016*(5), 1–37.

Pan, J.H., Ip, E. H., & Dubé, L. (2017). An alternative to post hoc model modification in confirmatory factor analysis: The Bayesian lasso. *Psychological Methods, 22*(4), 687–704.

Pan, J.H., Zhang, L.J., & Ip, E. H. *blcfa: Bayesian Lasso Confirmatory Factor Analysis.* Retrieved August 30, 2019, from https://github.com/zhanglj37/blcfa

Park, T., & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association, 103*(482), 681–686.

Rao, C. R. (1976). Estimation of Parameters in a Linear Model. *The Annals of Statistics, 4*(6), 1023–1037.

Richetin, J., Preti, E., Costantini, G., & De Panfilis, C. (2017). The centrality of affective instability and identity in Borderline Personality Disorder: Evidence from network analysis. *PLoS One, 12*(10), 1–14.

Rosenberg, M. D., Casey, B. J., & Holmes, A. J. (2018). Prediction complements explanation in understanding the developing brain. *Nature Communications, 9*(1), 1–13.

Scheidt, C. E., Hasenburg, A., Kunze, M., Waller, E., Pfeifer, R., Zimmermann, P., et al. (2012). Are individual differences of attachment predicting bereavement outcome after perinatal loss? A prospective cohort study. *Journal of Psychosomatic Research*, *73*, 375–382.

Schmid, N. S., Taylor, K. I., Foldi, N. S., Berres, M., & Monsch, A. U. (2013). Neuropsychological signs of Alzheimer's disease 8 years prior to diagnosis. *Journal of Alzheimer's Disease*, *34*, 537–546.

Serang, S., Jacobucci, R., Brimhall, K. C., & Grimm, K. J. (2017). Exploratory mediation analysis via regularization. *Structural Equation Modeling: A Multidisciplinary Journal*, *24.* 733–744.

Shi, G., Boerwinkle, E., Morrison, A. C., Gu, C. C., Chakravarti, A., & Rao, D. C. (2011). Mining gold dust under the genome wide significance level: A two-stage approach to analysis of GWAS. *Genet. Epidemiol*, *35,* 111–118.

Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science, 10*(6), 886–899

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *The Journal of Experimental Education, 70*(1), 80—93.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, 58*(1), 267—288.

Tibshirani, R., Friedman, J., Hastie, T., Narasimhan, B., Simon, N., & Qian, J. *glmnet: Lasso and Elastic -Net Regularized Generalized Linear Models.* Retrieved May 18, 2019, from https://www.rdocumentation.org/packages/glmnet/versions/2.0-18

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistic Society. B, 67*, 91—108.

Van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A system atic review of Bayesian articles in psychology: The last 25 years. *Psychology Methods, 22*(2), 217–239.

Waldmann, P., Mészáros, G., Gredler, B., Fuerst, C., & Sölkner, J. (2013). Corrigendum: evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in Genetics, 4*(4), 270.

Wilkinson, L. (1979). Tests of significance in stepwise regression. *Psychological Bulletin*, *86*, 168–174.

Wu, T.T., & Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics, 2*(1), 224–244.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science A Journal of the Association for Psychological Science*, *12*(6), 1100–1122.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68*(1), 49–67.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistic Society, 67,* 301–320.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of American Statistical Association, 101,* 1418–1429.

Zou, H., Hastie, T., & Tibshirani, R. (2007). On the “degrees of freedom” of the Lasso. *The Annals of Statistics, 35*, 2173–2192.

附录 1：Lasso 回归实例演示

为了验证传统 OLS 估计法容易出现过拟合的问题，展示 Lasso 回归的步骤和报告标准，促进 Lasso 回归的应用，本文将采用实例演示详细展示 Lasso 回归的分析流程，并对比传统估计方法。同时，实例分析还将纳入 Relaxed Lasso 方法。分析采用 R 软件，具体代码详见附录 2。

数据来源于 395 名葡萄牙中学生(Cortez & Silva, 2008)，数据中包含了 11 个连续变量：(1) 年龄(age)，(2) 家庭关系质量(famrel)，(3) 放学后空闲时间(freetime)，(4) 和朋友出去玩的频率(goout)，(5) 工作日饮酒频率(dalc)，(6) 周末饮酒频率(walc)，(7) 自评健康状况(health)，(8) 缺课次数(absences)，(9) 学生第一次数学测验成绩(G1)，(10) 中期测验成绩(G2)和 (11) 期末测验成绩(G3)。其中期末测验成绩为因变量，本研究将探究能够有效预测数学期末测验成绩的因素。相关分析结果显示，学生第一次数学测验成绩、中期测验成绩与期末测验成绩之间存在较强的正相关。

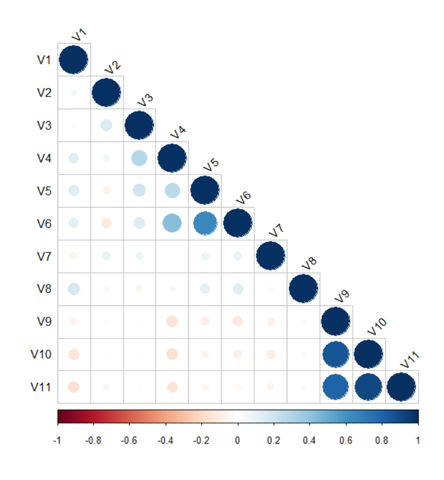


图 1 变量间相关图

注：红色系代表负相关，蓝色系代表正相关，颜色越深代表相关值越大。

在 Lasso 回归中，首先采用 10 重交叉验证方法选择合适的惩罚项 λ。这一方法可以通过 R 软件中的 glmnet 包(Friedman, Hastie, & Tibshirani, 2010)实现。值得注意的是，为了保证每次交叉验证分析得到的 λ 结果一致，需要采用 set.seed()函数设定随机数种子，否则每次

分析的结果会存在微小差异。

结果显示最小化均方误差(Mean Square Error, MSE)的 λ 为 0.043，λ + 1se 为 0.776。图 2 呈现了随着 log(lambda)的增加 MSE 值的变化。当 λ 对复杂模型的惩罚力度增大时，MSE 同样会增大，而惩罚项的增大最终会导致所有系数压缩到 0，此时 MSE 值最大。

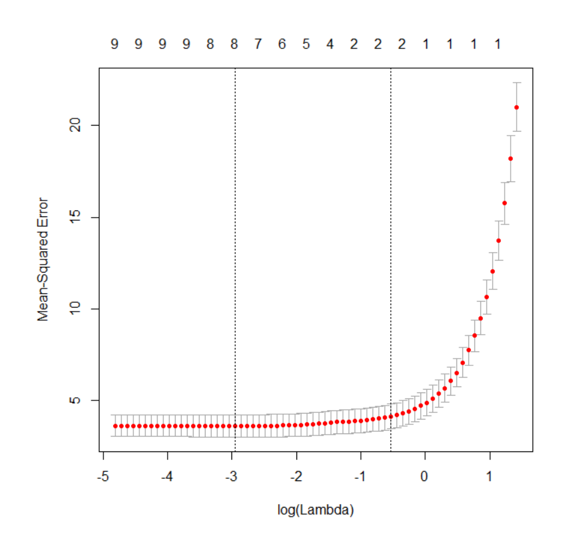


图 2 十重交叉验证结果

注：图中两条竖线分别代表最小化 MSE 的 λ 值和 λ + 1se 值

图 3 呈现了随着 log(lambda) 的增加，标准化回归系数被压缩的情况，可以看到的是，随着惩罚力度的增大，标准化系数最终全部会被压缩到 0。而在 λ 值为 0.776 处，有两个系数不为 0。根据输出结果，G1(学生第一次数学测验成绩)和 G2(学生中期数学测验成绩)两个预测因素被保留下来。

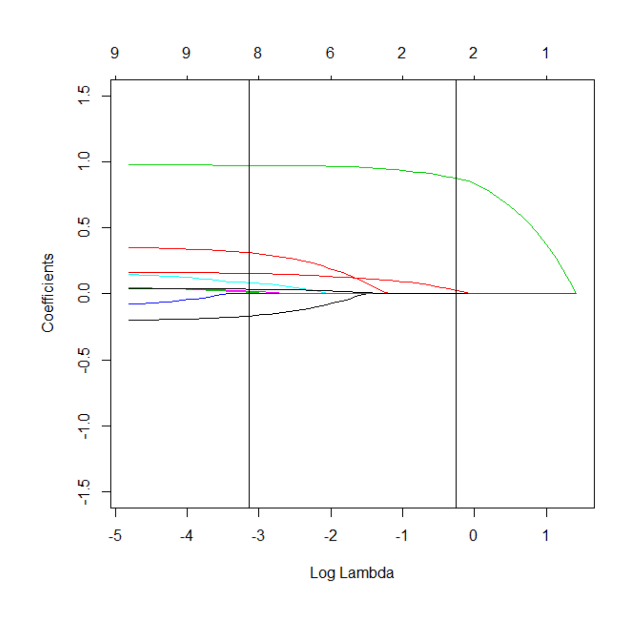


图 3 惩罚项对系数的压缩结果

此外，Lasso 回归中可以通过 covTest 软件包(Lockhart et al., 2014)计算参数估计的 p 值，进一步计算 p 值发现，同样只有 G1 和 G2 变量通过了显著性检验(表 1)。

而在 OLS 估计中，共发现了年龄、家庭关系质量、缺课次数，第一次测验成绩和期中成绩五个变量可以显著预测期末数学成绩(表 1)。但是结果显示缺课次数正向预测期末数学成绩，即学生缺勤次数越多，期末成绩越高($b = 0.042, p = 0.001$)，这显然和常识相悖。而相关分析也显示缺课次数和期末成绩间未发现显著相关($r = 0.034$, $p = 0.497$)。而 OLS 回归分析得到的显著结果可能是由于样本量和观察指标数的比率较低($n / p = 3.95$)，模型发生了过拟合现象，即模型在最小化结果变量的预测值和观测值的差异时，错误地学习到了不存在的规律。此外，和 Lasso 回归相比，OLS 额外发现的另外两个显著的预测变量和期末成绩的相关值较弱(图 1)。其中年龄和期末数学成绩显著负相关($r = -0.162, p = 0.001$)，而家庭关系质量和期末数学成绩未发现显著相关($r = 0.051, p = 0.309$)。

进一步进行 Relaxed Lasso 分析，即采用 Lasso 回归选择出的 G1 和 G2 变量与期末数学成绩建立 OLS 回归模型。结果发现与传统的 OLS 估计相比，Relaxed Lasso 回归的 R^2、校正后 R^2 及均方误差)均相差不大。即 Relaxed Lasso 回归仅采用两个预测变量就基本达到了 OLS 回归采用 5 个变量所获得的预测能力。

表 1　Lasso、 OLS、Relaxed Lasso 回归结果

预测变量	系数估计值(p 值)		
	OLS	Lasso	Relaxed Lasso
age	-0.206(0.009)**	-(0.072)	-
famrel	0.36(0.001)**	-(0.699)	-
freetime	0.058(0.57)	-(0.913)	-
gout	-0.014(0.891)	-(0.981)	-
dalc	-0.108(0.448)	-(0.646)	-
walc	0.17(0.105)	-(0.294)	-
health	0.046(0.509)	-(0.899)	-
absences	0.042(0.001)**	-(0.089)	-
G1	0.164(0.003)**	0.057(0.005)**	0.153(0.007)**
G2	0.977(<0.001)***	0.903(<0.001)***	0.987(<0.001)***

R^2	0.835	-	0.822
adjusted R^2	0.831	-	0.821
Mean Square Error	3.446	-	3.723

注：**代表 p 小于 0.01，***代表 p 小于 0.001。

从上述分析中可以看出，OLS 回归所选择的预测变量可能是不可靠且冗余的。一方面在本研究中 OLS 回归所选择的预测变量和因变量间相关很弱，另一方面，增加的三个预测变量并不能很好地提升对因变量的解释力，R^2 和校正后 R^2 的值都和仅采用两个预测变量的回归模型相接近。此外，Relaxed Lasso 方法也避免了 Lasso 方法在压缩不重要的系数的同时对非零系数(G1, G2)的压缩。

值得注意的是，Lasso 回归并不总是会获得更简洁的预测变量集，它的目的是采用较少的预测变量获得较高的预测能力。这尤其体现在样本量较少时，OLS 回归所使用的假设检验为了控制一类错误率，通常会获得较高的标准误估计，检验力较低，而 Lasso 回归在此时则更易于获得更高的检验力和预测能力。

附录 2：Lasso 回归实例代码

```
student <- read.table("mat_2.txt",sep="\t",header=FALSE)

IV<-(student[,1:10])

IV1=scale(IV,FALSE,FALSE) ## 不对自变量进行标准化处理


## 十重交叉验证

install.packages('glmnet')

library(glmnet)

set.seed(1222)  ## 设定随机数种子，保证每次运行十重交叉验证的结果一样

Lambda=cv.glmnet(IV1,student[,11])


## lasso 回归结果

coef(Lambda, s=Lambda$lambda.1se)


## 绘图

plot(Lambda)  ## 横坐标为 lambda，纵坐标为均方误差 MSE

savePlot(filename = "lambda", type ="png", device = dev.cur(),

        restoreConsole = TRUE)


RegCoef=glmnet(IV1,student[,11],family = "gaussian",alpha = 1)

plot(RegCoef, xvar="lambda",ylim=c(-1.5,1.5), lwd=1.8 )

        ## 横坐标为 lambda，纵坐标为系数估计值

abline(v=log(Lambda$lambda.1se))

abline(v=log(Lambda$lambda.min))

savePlot(filename = "loglambda", type ="png", device =

dev.cur(),

        restoreConsole = TRUE)
```

```
## 采用 covTest 包计算 p 值
library('devtools')
install_github('cran/covTest')
## coveTest 软件包目前无法从 CRAN 上下载，因此采用 devtools 软件包从
github 上下载
library(covTest)

IV<-student[,1:10]
df=nrow(IV)-1
IV2=scale(IV,TRUE,TRUE)/sqrt(df)   ## 标准化自变量

LarsCoef=lars(IV2,student[,11])
covTest(LarsCoef,IV2,student[,11]) ## 计算 p 值
```

Lasso regression: From explanation to prediction

ZHANG Lijin[1]; WEI Xiayan[2]; LU Jiaqi[2]; PAN Junhao[1]

([1] *Department of Psychology, Sun Yat-sen University, Guangzhou 510006, China*)

([2] *Department of Psychology and Behavioral Sciences, Zhejiang University, Hangzhou 310028, China*)

Abstract: Psychological researches focus on describing, explaining and predicting behavior, and having a good understanding of the association between variables is an essential part of this process. Regression analysis, a method to evaluate the relationship between variables, is widely used in psychological studies. However, due to its highly focus on the interpretation of sample data, the traditional ordinary least squares regression has several drawbacks, such as over-fitting problem and limitation on dealing with multicollinearity, which may undermine the generalizability of the model. These drawbacks have an inevitable influence on the promotion and prediction of the model conclusion.

With the rapid development of methodology, Least absolute shrinkage and selection operator (Lasso) regression has been emerged to better compensate for the limitations of traditional methods. By introducing a penalty term in the model and shrinking the regression coefficients to zero, Lasso regression can achieve a higher accuracy of model prediction and model generalizability with the cost of a certain estimation bias. Besides, Lasso regression can also effectively deal with the multicollinearity problem. Therefore, it has been widely used in medicine, economics, neuroscience and other fields.

In psychology, due to the limitations of computer computing power, researchers used to mainly rely on hypothesis testing to understand the association among variables to verify theories. Now, with the rapid development of machine learning, a shift from focusing on interpretation of the regression coefficients to improving the prediction of the model has emerged and become more and more important. Therefore, based on fundamental theories and real data analysis, the aim of this paper is to introduce the principles, implementation steps and advantages of the Lasso regression. With the help of statistic science, it is promising that more and more applied researchers will be called upon to focus on the emerging statistical tools to promote the development of psychology.

Key words: regression, regularization, Lasso, prediction